



King's Research Portal

DOI:

[10.1021/ac4024878](https://doi.org/10.1021/ac4024878)

Document Version

Peer reviewed version

[Link to publication record in King's Research Portal](#)

Citation for published version (APA):

Miller, T. H., Musenga, A., Cowan, D. A., & Barron, L. P. (2013). Prediction of chromatographic retention time in high-resolution anti-doping screening data using artificial neural networks. *Analytical Chemistry*, 85(21), 10330-10337. [N/A]. <https://doi.org/10.1021/ac4024878>

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

This document is confidential and is proprietary to the American Chemical Society and its authors. Do not copy or disclose without written permission. If you have received this item in error, notify the sender and delete all copies.

PREDICTION OF CHROMATOGRAPHIC RETENTION TIME IN HIGH RESOLUTION ANTI-DOPING SCREENING DATA USING ARTIFICIAL NEURAL NETWORKS

Journal:	<i>Analytical Chemistry</i>
Manuscript ID:	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Miller, Thomas; King's College London, Analytical & Environmental Sciences Musenga, Alessandro; King's College London, Analytical & Environmental Sciences Cowan, David; King's College, Forensic Science and Drug Monitoring Barron, Leon; Kings collodge London, Forensic & Analytical Sciences

SCHOLARONE™
Manuscripts

1 PREDICTION OF CHROMATOGRAPHIC
2 RETENTION TIME IN HIGH RESOLUTION
3 ANTI-DOPING SCREENING DATA USING
4 ARTIFICIAL NEURAL NETWORKS

5
6 *Thomas H. Miller, Alessandro Musenga, David A. Cowan, Leon P. Barron**

7 Analytical & Environmental Sciences Division, School of Biomedical Sciences,
8 King's College London, 150 Stamford Street, London, SE1 9NH, United Kingdom.

9
10 *Corresponding author

11 Email: leon.barron@kcl.ac.uk

12 Tel: +44 20 7848 3842

13 Fax: +44 20 7848 4980

14
15
16 **Keywords:** artificial neural networks; anti-doping; high resolution mass
17 spectrometry; ultra-high pressure liquid chromatography; retention time

Abstract

The computational generation of gradient retention time data for retrospective detection of suspected sports doping species in post-analysis human urine sample data is presented herein. Retention data for a selection of 86 compounds included in the London 2012 Olympic and Paralympic Games drug testing schedule were used to train, verify and test a range of computational models for this purpose. Spiked urine samples were analysed using solid phase extraction followed by ultra-high pressure gradient liquid chromatography coupled to electrospray ionisation high-resolution mass spectrometry. Most analyte retention times varied ≤ 0.2 min over the relatively short runtime of 10 min. Predicted retention times were within 0.5 min of experimental values for 12 out of 15 blind test compounds (largest error: 0.97 min). Minimising the variance in predictive ability across replicate networks of identical architecture is presented for the first time along with a quantitative discussion of the contribution of each selected molecular descriptor towards the overall predicted value. The performance of neural computing predictions for isobaric compound retention time is also discussed. To the authors' knowledge, this work is the first to present the development and application of neural networks to the prediction of gradient retention time in archived urine analysis sample data in the field of anti-doping.

1 Introduction

Drug testing in sport is becoming more and more challenging, since cheats rapidly switch to new or designer drugs to avoid being caught by anti-doping tests. Because of this, in 2011 the World Anti-Doping Agency introduced a new section in the Prohibited List to specify that any substance not approved for human therapeutic use is prohibited.¹ Unfortunately however, there is normally a delay between new substances being used by cheats and the screening methods being implemented to detect them.

New technologies in analytical chemistry now exist which offer a much more flexible and universal approach compared to targeted-type screening analyses. In particular, liquid chromatography with full data capture high resolution mass spectrometry (LC-HRMS) is becoming very popular for drug screening, since it generally permits the addition of new analytes to a screening method with little/no modification to the method. One of the most promising features is the possibility for retrospective analysis of post-acquisition datasets.^{2,3} The flexibility of the HRMS approach was exploited in our laboratory during the London 2012 Olympic and Paralympic Games, where an LC-HRMS screen was employed for the detection of nearly 200 analytes within a single assay.⁴ When permitted, in accordance with the International Standard for Laboratories,⁵ all sample data acquired by this technique could be re-processed and re-evaluated (for example, where intelligence becomes available about the potential use of new substances). However, several challenges still exist for retrospective detection of drug species in both new and archived samples. Analytical reference standards are not always available, especially for new or analogous compounds which may be synthesised relatively easily and with sufficient efficacy for

1 practical use by athletes. Moreover, the exact operating conditions of analytical
2 equipment may not be reproducible at a later date when a repeat analysis is required.
3 For example, chromatograms generated from separate analyses even using the same
4 columns may differ due to phase ageing and loss of performance. Therefore, more
5 effort is required to interpret the original data.

6
7 Predictive computing techniques, such as artificial neural networks (ANNs), can
8 interpret underlying trends in complex datasets by learning from case examples. There
9 are several types of ANN, but the feed-forward multi-layer type is commonly the
10 most used across a range of scientific applications such as analytical chemistry,
11 environmental science and molecular biology.⁶⁻¹⁰ By interconnecting a set of input
12 data with a series of hidden layer neurons, statistical weights and biases between them
13 are systematically optimised towards producing a minimised error overall output. In
14 the training phase, the ANN requires a known true value as a comparator and once an
15 acceptable number of training cycles (or epochs) is determined, the optimised ANN
16 can be used to predict the same output where experimentally-derived data is
17 unavailable (i.e. a blind test). In this case, information about a set of analytes or
18 chromatographic system (as inputs) and the experimentally-derived retention data for
19 standards within spiked, matrix-matched samples (as an output) could provide a
20 useful body of information for an ANN to predict retention time retrospectively under
21 the conditions observed at the time.

22 Research to characterise retention behaviour has used linear solvation energy
23 relationships (LSERs). These use solute descriptors and linear regression modelling to
24 predict a retention parameter for a particular set of compounds. Both algorithm-based
25 and predictive ANN approaches have incorporated the use of LSERs as inputs to

1 estimate retention (usually as retention factor, k) across a number of chromatographic
2 modes,¹¹⁻¹⁴ column formats,¹⁵⁻¹⁷ for method optimisation purposes,¹⁸⁻²² to predict
3 retention behaviour across column formats, or to estimate the retention of
4 unknowns.²³⁻²⁵ Each LSER solute descriptor must be calculated and each coefficient
5 is usually obtained via measured retention data for a set of representative compounds.
6 Thus, the time to produce an accurate LSER model can be significant (as well as
7 especially challenging for ionisable compounds in particular^{15,18}). As LSER
8 coefficients for retention modelling also depend on observed retention on a stationary
9 phase, factors such as the age of the phase may lead to observable differences and
10 lower accuracy over time. Furthermore, the exact composition of a changing mobile
11 phase under gradient conditions is very difficult to characterise and requires the
12 inclusion of extra variables such as purity and individual system dwell time for
13 example. Whilst the investigation of LSERs in retention modelling advances, ANN-
14 based predictive approaches using alternative descriptors may therefore hold promise
15 for more immediate application.

16 The aim of this work was to investigate the use of ANNs for the prediction of
17 retention time in archived urine analysis data from the London 2012 Olympic and
18 Paralympic Games using a method employing ultra-high pressure liquid
19 chromatography coupled to HRMS. In particular, the selection of alternative
20 molecular descriptor data to those used in LSERs is presented, along with a study of
21 the ANN dependency on these descriptors and the variance across replicated ANNs.
22 Lastly, this work aimed to assess whether ANNs could be used to discriminate
23 structurally similar and/or isobaric species.

24

1 Experimental

2 *Reagents*

3 For a detailed account of all reagents used in the experimental please see the
4 accompanying supplementary information (SI), Table S2. For those compounds where
5 no certificate of analysis was available, a solution of each compound was infused
6 separately into the HRMS and full scan data acquired without any applied
7 collisionally induced dissociation. The internal calibrator was caffeine. Identity was
8 considered confirmed where m/z values within 2 ppm of the theoretical value were
9 achieved along with a consistent fragmentation pattern.

10 *Urine sample preparation*

11 Drug-free urine samples were collected from 15 healthy volunteers, anonymised,
12 divided into 1 mL aliquots and stored at -10 °C until analysis. Each aliquot was then
13 spiked with all reference compounds. Following addition of internal standards and
14 glucuronide hydrolysis reagents (see SI, 2.0), formic acid was added to the samples.
15 This solution was then extracted on Bond-Elut Plexa PCX (60 mg, 3 mL barrel) solid
16 phase extraction (SPE) cartridges (Agilent Technologies, Lake Forest, CA).
17 Cartridges were pre-conditioned with 0.5 mL of methanol and 0.5 mL of 2 % v/v
18 formic acid. After loading, cartridges were washed with 1 mL of each of 2 % formic
19 acid, water and 20:80 v/v methanol:water. Elution was performed in 3 mL 3 % (v/v)
20 ammonium hydroxide in methanol:acetonitrile (50:50, v/v). Extracts were dried under
21 nitrogen at 60 °C and then reconstituted to 100 μ L of 0.3 % v/v formic acid in 95:5
22 v/v water:acetonitrile.

23

Instrumental conditions

For all separations, a Waters Acquity UPLC[®] ultra-high pressure liquid chromatographic system was used (Waters, Milford, MA, USA). Separations were performed on a Waters Acquity BEH-C₁₈ column (2.1 × 50 mm, 1.7 μm). Gradient elution was performed as follows: 95:5 water:acetonitrile (both in 0.3 % formic acid) for 0.5 min; then to 80:20 for 3.0 min; a linear ramp to 75:25 for 2.0 min; to 43:67 for 1.5 min; and finally to 10:90 for 1 min. Total time for the gradient elution was 8 min, followed by 2 min for re-equilibration. Separations were performed at 0.3 mL/min and at 30 °C throughout. The injection volume was 10 μL. The syringe and injector were washed sequentially with 0.2 % formic acid in acetonitrile/water (30:70, v/v) and methanol/acetonitrile (90:10, v/v) up to 5 times each before every run to avoid carryover.

Detection was performed using fast polarity switching high resolution mass spectrometry on an Exactive instrument (Thermo Fisher Scientific, San Jose, USA) equipped with a heated electrospray ionisation (HESI-II) source. Enhanced resolution mode was employed at 25,000 FWHM resolution. Three events occurred during each acquisition cycle by performing a full scan both in positive and negative ionisation mode (both with disabled CID) followed by a full scan in positive ionisation mode only with CID (HCD collision energy 30 eV). For a full description of other HRMS conditions, see the SI.

Molecular descriptors, neural network prediction models and architectures

Eighteen molecular descriptors were generated for each compound in the optimised network. Of these, 15 descriptors were computed using Parameter Client freeware

(Virtual Computational Chemistry Laboratory, Munich, Germany) using canonical simplified molecular-input line entry system strings (SMILES). Descriptors including pK_a , *Ghose-Crippen and Moriguchi log P* (AlogP or MlogP), number of double bonds (nDB), number of four to nine-membered rings (nR04-nR09) and number of carbon or oxygen atoms (nC or nO) were used for network optimisation. A full list of other descriptors is detailed in Table S1 and Figure 3. To investigate the use of predicted pK_a values, Percepta PhysChem Profiler (ACD Labs, ON, Canada) software was used. All predicted retention times (t_r^P) were performed using licenced ANN software (Trajan Software Ltd., Lincolnshire, UK) and compared with experimentally determined retention time (t_r^E). A selection of ANN types and architectures were investigated including linear, radial basis function (RBF), probabilistic neural networks (PNN) and multi-layer perceptrons (MLPs).

Results and Discussion

Experimentally determined chromatographic retention time and reproducibility

This screening method was used during the 2012 Olympic Games for the detection of nearly 200 compounds in our ISO 17025 accredited laboratory. All of these species were detectable at concentrations corresponding to 50 % of the WADA minimum required performance level²⁶ in force at that time in spiked, drug-free urine. On average, peak widths at baseline were of the order of ~0.15 min with some notable exceptions. For example, peak widths for morphine and etilefrine were 0.3 and 0.4 min respectively, with some variability between urine samples. In order to examine the potential usefulness of computational tools for the retrospective generation of retention times in archived data, the reliability of generated

1 chromatographic retention data was first evaluated. An in-depth evaluation of other
2 method validation experiments including selectivity, sensitivity and limits of detection
3 is given elsewhere.⁴ The evaluation of retention time reproducibility was performed
4 using 15 different urine samples analysed in groups of 5 with groups being analysed
5 on different days. Rather than expressing the reproducibility as a relative standard
6 deviation, the maximum variation in retention time was preferred, since the aim was
7 to establish a suitable detection window. The maximum within-day t_r^E variability was
8 0.22, 0.23 and 0.32 min for the 3 groups of 5 urine samples respectively. Considering
9 all 15 urine samples together (3 groups, 3 different days), the maximum retention time
10 variability was 0.35 min (i.e. ± 0.175 min). Data analyses during screening were
11 normally performed by reviewing extracted ion chromatograms with an m/z window
12 of ± 5 ppm and a retention time window of ± 0.5 min. Athlete urine samples were
13 always run using a “bracketed” approach to take into account phase ageing and to
14 ensure system performance throughout the entire batch acquisition.

15 *Selection of molecular descriptors, network type and architecture*

16 By generating SMILES strings for each compound, an initial set of >200 diverse
17 molecular descriptors were generated. Whilst the instrumental screening method
18 incorporated a larger number of compounds in practice, the number studied here was
19 limited to those compounds with literature reported, experimentally-derived pK_a
20 values (as this descriptor was not initially available within the Parameter Client).
21 Given that the majority of these compounds were fully/partly ionised in the mobile
22 phase, this descriptor was considered important for the potential prediction of
23 retention times. Multiple network types and architectures were examined. On the
24 whole, it was observed that use of a larger numbers of inputs did not offer advantages

1 in terms of prediction accuracy and a much smaller set was ultimately more practical
2 and yielded better correlations. As a result, some descriptors were not included in the
3 model which might have otherwise been expected under reversed-phase
4 chromatography conditions such as analyte topological surface area, number of atoms
5 containing lone pairs of electrons (e.g. oxygen or nitrogen atoms), or molecular
6 weight.

7
8 Whether using a large or small input descriptor set, a linear-type network offered no
9 detectable correlation (R^2 and slope both <0.1 in all cases) between t_r^E and t_r^P and
10 irrespective of whether inputs were used for training, verification or blind testing. As
11 such, this network type was removed from consideration. Alternative network types
12 on the other hand offered better performance and the best correlation coefficients
13 achieved for each type ranged from $R^2=0.86$ (RBF) to 0.98 (4-layer MLP) with the
14 slope (m) ranging from 0.97 (3-layer MLP) to 1.03 (PNN), again for all 86
15 compounds. The best network shown in Figure 1(a) was a feed-forward, back
16 propagation-type MLP with 2 hidden layers of 5 and 4 nodes respectively. This
17 18:5:4:1 MLP architecture was trained using 18 molecular descriptors using 61
18 compounds. Verification was performed with 10 compounds and training of the
19 network was ceased when a minimum for residual error was observed (2,000 epochs
20 here). The network was “blind” tested using 15 additional compounds displaying
21 structural variance and a spread of input data and t_r^E data across the 8 min runtime.
22 Data for t_r^E was not used by the ANN for test compounds (see Table S2) and
23 performance was observed after training and verification. When the dataset was
24 divided into training, verification and blind testing data, their respective m and R^2
25 values remained satisfactory and were also better than the corresponding plots for

PNN, RBF and 3-layer MLP networks. By examining the residual errors in the entire dataset (Figure 1(b)), it was observed that retention time could be predicted within 0.5 min of t_r^E for 80 out of 86 compounds (~93 %). This window was initially selected to cover the window employed during experimental screening outlined above. In particular, t_r^P for 97 %, 90 % and 80 % of the training, verification and test compounds respectively lay within this window. In fact for the test set alone, this level of accuracy was maintained even to ± 0.3 min of t_r^E . Therefore, these results show that this type of network offered promising predictive ability for most compounds in a relatively rapid gradient separation. The largest residual errors were observed for MDA, clenbuterol and *p*-methylanphetamine at -0.947 min (used within the test set), -0.721 min (in verification set) and -0.704 min (in training set) respectively. Therefore, all species could be predicted within 1 min of t_r^E . However, it should be noted that this ANN was specifically tailored for this chromatographic system and method. If applied to other separation systems and modes, it is likely that similar descriptor and ANN optimisation experiments would be necessary.

Consistency in t_r^P using replicate networks of identical type and architecture

One of the main issues with ANNs is that training follows a random process towards minimising the overall error. However at larger numbers of training epochs, networks may become over-trained and whilst training set errors may be minimised, the verification set error can increase (as it is not used by the software itself during the training process). The user often manually defines where training ceases by observing where the combined verification set residual error is at its lowest. In some cases this verification error is lowest at lower numbers of training epochs, but where training errors may remain high. Therefore, the variability in network connection weightings

1 may be very different from network to network (even with identical architectures) and
2 a balance is required potentially towards achieving some consistency. In this case,
3 replicate networks (n=10) using the optimised architecture were built to examine
4 consistency and the range of t_r^P values for the verification and test sets were plotted as
5 Figure 2. Also shown alongside each t_r^P entry in Figure 2 is the t_r^E variance of the
6 analytical method for comparison purposes. By stopping training between 2,000-
7 3,900 epochs in each case (where verification and training error combined were
8 lowest), it was found that relatively consistent input weightings were achieved and t_r^P
9 performance could be maintained. To our knowledge, this has not been shown
10 previously in predictive chromatographic retention studies. However, and despite their
11 accuracies and consistencies, t_r^E data for 6/15 blind test set compounds still lay
12 outside the range of t_r^P values generated by all 10 networks. Of these, and of particular
13 interest, were three structurally-related species: dimethylamphetamine, MDA and
14 phentermine. As synthetic compounds of this general type often only have slight
15 variations in their molecular structure, this approach could be very useful to predict
16 retention time for an array of related species where a reference standard is
17 unavailable. The average t_r^P errors across all 10 networks for these three compounds
18 were +0.97, -0.62 and -0.24 min (t_r^E =2.76, 2.50 and 2.91 min) respectively. The
19 lowest respective errors in t_r^P achieved by any one network were +0.59, -0.57 and -
20 0.07 min. By comparison, and upon inspection of the average t_r^P error for other related
21 species within the training set, compounds such as MDMA, benzphetamine,
22 fenfluramine and methamphetamine all performed reasonably well (average absolute
23 error of 0.08-0.18 min from t_r^E for all 10 networks). However, and as noted earlier,
24 average t_r^P error for *p*-methamphetamine was the worst case in the training set at an
25 average of -0.44 min across all networks. In the verification set, only one related

1 compound, amphetamine, yielded 0.13 min error in t_r^P ($t_r^E=2.37$ min). Therefore,
2 these figures suggest that dimethylamphetamine and MDA may be unusual cases and
3 that ANN predictive ability for unknowns may not, on average, be poorer for these
4 types of compound. With regard to the general test set, all verification and test
5 compound retention times could be predicted within 1 min of t_r^E (with
6 average±standard deviation of t_r^P error for all 25 compounds = 0.3 ± 0.2 min for all 10
7 networks). In comparison to the maximum measured experimental variance above
8 (0.35 min), the potential for using multiple networks to predict retention time shows
9 promise. Therefore, the window for ANN accuracy could be set at ±0.5 min in total
10 (again, also comparable to the experimental retention time window used in practice
11 during the Olympic & Paralympic Games). Each network required a few minutes to
12 build, train, verify and test. Therefore, this *in silico* approach could represent a
13 significant saving in terms of time, effort and cost when used in combination with a
14 semi-targeted post data acquisition analysis of an athlete urine sample. Samples from
15 the Olympic Games are stored for a total of 8 years in case re-analysis is required at a
16 later date. However, this is not the case for most other sporting events where negative
17 urine samples are discarded after 3 months unless probable cause for prolonged
18 storage is justified. As several isobaric compounds can exist, HRMS data alone may
19 be considered insufficient for this purpose. Therefore, the benefits of ANN predictions
20 could aid in the investigation of doping before analysis, before the reporting stage and
21 after the reporting stage. When used in combination with HRMS, this gives added
22 value especially in directing which analytical reference materials are to be
23 synthesised.

24 *Molecular descriptor contribution towards t_r^P and variance across replicate networks*

Using the optimised MLP network, the contribution of each descriptor towards the generation of t_r^P was calculated for a single compound, A , by determining the absolute deviation, $D_{A,i}$, arising from sequential removal of one descriptor, i , at a time.²⁷ Data were inputted into Equation 2 below to calculate the overall variance, V_A , arising for that compound. Values for V_A for each compound (denoted $A=1-61$) were then used to determine the overall percentage change in accuracy of predicted retention time for all compounds, $\Delta t_r^P(\%)$, for the removal of that descriptor according to Equation 3.

$$V_A = \frac{(D_A)_i}{\sum_{i=1-18} (D_A)_i} \quad (2)$$

$$\Delta t_r^P(\%) = \frac{100.V_A}{\sum_{A=1-61} V_A} \quad (3)$$

The contribution of each descriptor towards t_r^P was different. Using the optimised network, the top 5 contributing molecular descriptors (in decreasing order) to t_r^P for all compounds using the best network were nC (17.5 % change in t_r^P when removed), nR09 (10.0 % change), nO (9.9 % change), nDB (8.8 % change) and ALogP (8.2 % change). Overall, some of these Δt_r^P values were unsurprising given that silica-C₁₈ was used as a stationary phase. For experimentally-derived descriptors however, removal of pK_a only resulted in a change of 3.8 %, but subsequent removal of AlogD and MlogD resulted in Δt_r^P of ≤ 3 % each. As expected, descriptors such as the number nR04 having little or no incidence resulted in small Δt_r^P when removed. Examination of the data relative to t_r^E revealed that predictions were on the whole worse with the removal of selected descriptors.

To further understand the variability across replicate networks, this experiment was repeated for 9 additional replicate networks. The contribution and variance of each descriptor is shown as Figure 3. Taking the mean values, the top 5 contributing

1 molecular descriptors were the same as reported above, albeit in a slightly different
2 order (from highest to lowest: nC, nDB, nO, AlogP and nR09). This indicated that
3 ANNs could be replicated with acceptable consistency in their operation.
4 Interestingly, the largest measured variance in t_r^P across replicate ANNs for removal a
5 single descriptor was also observed as the most highly contributing descriptor, nC, at
6 ~11 %. Removal of descriptors which resulted in the lowest % t_r^P variance were also
7 those which were used least by ANNs during prediction. However for the rest of the
8 descriptor set, no trend was apparent to conclude that a more heavily used input would
9 result in a higher variance when omitted from replicate networks. Furthermore, it was
10 observed again that AlogP was repeatedly used in preference to MlogP in t_r
11 predictions. MlogP has been shown previously to be a less accurate descriptor in
12 comparison to other available computational models and in respect to experimentally-
13 derived logP.²⁸ Here, it was observed that MlogP data offered poorer discrimination
14 between some species and especially those of isobaric mass. It was also apparent that
15 this had a consequential effect on the use of AlogD over MlogD as a preferred input
16 descriptor of the two. However, as a null % change was not observed by removing
17 MlogP from the dataset in any network, it was still used to some degree by ANNs.

18 *Substitution of experimentally-derived pK_a with predicted pK_a*

19 In practice, retention on reversed-phase media is likely to be heavily dependent on
20 analyte logP and pK_a (and subsequently logD). The molecular description software
21 used here could generate predicted analyte logP, but not pK_a . The initial use of
22 experimentally-derived pK_a data from the literature overcame this problem and
23 resulted in satisfactory t_r^P accuracy as shown earlier. However, the availability of (or
24 indeed the ability to generate) experimental pK_a significantly limits the usefulness of

1 this ANN approach if it is to be used in post-acquisition data mining. The use of a
2 separate pK_a prediction software package was therefore investigated as a possible
3 alternative. In an initial experiment, experimentally-derived pK_a data were removed
4 from the input dataset and replaced directly with predicted values into the pre-trained
5 ANN. With this substitution, a very slight reduction in t_r^P accuracy was observed, but
6 it still remained acceptable (not shown). The network in Figure 1(a) generated an
7 overall t_r^P / t_r^E correlation with a slope of 1.00 ($R^2 = 0.98$) compared with 0.95
8 ($R^2=0.98$) in this case across all 86 analytes. This obviously indicated some minor
9 discrepancies between numerical values in these datasets. To further understand the
10 importance of predicted versus experimentally-derived pK_a as ANN inputs, a new
11 network of identical architecture was retrained using computationally-derived pK_a
12 values to observe whether prediction accuracy could be restored. From this, it was
13 seen that t_r^P accuracy was marginally worse using computed pK_a (Figure S3), but still
14 remained acceptable overall (overall slope=0.94; $R^2=0.95$). In previous work,
15 Livingstone also showed that predicted pK_a data can display inaccuracies up to 1.5 log
16 units (using a larger compound set than here).²⁸ Thus, in ANN predictions of test
17 compound retention time, experimentally-derived data should be used where possible,
18 but predicted values could potentially be substituted with awareness that accuracy
19 may be slightly affected.

20 *Performance of ANNs for isobaric and isomeric compounds*

21 If the ANN approach is to work in any realistic scenario with HRMS detection, then
22 t_r^P accuracy should be high for compounds which are both isobaric and which display
23 high degrees of structural similarity. From the list of compounds used in this study,
24 several such cases existed (Table 1). Nadolol and metipranolol, both non-selective

beta-blockers, display identical mass and molecular formula, but differ slightly in their molecular structure and logP. Experimentally, this translated to a two-fold difference in their retention times on this C₁₈ phase. In this case, the optimised ANN performed well by predicting retention within 0.4 min for both compounds despite their high structural and pK_a similarity. Differences in logP could have accounted for this performance, as it was used here in the calculation of logD. Hydrophilic factor was calculated based on the number of atoms in a molecule, the number of carbon atoms and the number of hydrophilic groups (such as –OH or –NH for example). Unsaturation indices are calculated based on the number of atoms containing proton deficiencies (e.g. aromatic rings or multiple bonds). Therefore, these latter two descriptors could be considered loosely correlated with logP on a fundamental level. Therefore, it might be somewhat unsurprising that this resulted in higher accuracy in this case given the nature of the separation mode.

In a second similar case involving the isobaric compounds, phentermine and methamphetamine t_r^P accuracy remained high (± 0.22 min from t_r^E), but the order of elution was incorrectly predicted. Furthermore, t_r^P for a third isobaric compound, *p*-methamphetamine was predicted as similar to that of phentermine and methamphetamine, but was less accurate to its t_r^E at -0.7 min. All three of these compounds had identical MlogP data, but only slightly differing AlogP and pK_a values. For the majority of the other 13 compounds eluting between methamphetamine and *p*-methamphetamine and where descriptor diversity was more pronounced, this was less of a problem and t_r^P inaccuracy remained <0.5 min.

As a third case example, a set of diastereoisomers were chosen: ephedrine and pseudoephedrine. With identical mass, molecular formula, atomic connections, AlogP and MlogP, it was somewhat unsurprising that they remained experimentally unresolved on this stationary phase. Whilst their pK_a values were 0.3 units apart, and therefore a little further apart than of those species in the previous example, this had little effect on t_r^P (within 0.15 min of t_r^E for both species). Whilst further work on the use of ANN using resolved diastereoisomers is required, it should be noted that the SMILES strings used here employs a string of characters that specified the structure of a selected molecule including its spatial arrangements. Optical isomerism in SMILES strings is designated by the '@' symbol. Where a single '@' is used, this represents the substituent groups giving rise to the chirality being positioned anti-clockwise around the chiral atom whereas two symbols ('@@') indicates clockwise orientation of the groups around the chiral atom. Therefore, as this may lead to a difference in molecular description, then this could be used by the ANN to differentiate such species. Naturally, this would also depend on the importance of the molecular descriptor itself to the generation of t_r^P (similar to Figure 3).

Conclusion

This work showed that ANNs could be used to predict chromatographic retention times for 93 % of all selected doping-related compounds to within 0.5 min of their true value and to within 1 min for all other compounds. All compounds were detected in a urine extract matrix and were separated under gradient elution conditions. When applied to the prediction of unknowns alone, the same level of accuracy was maintained. Variance across replicate networks of identical architecture revealed that descriptors were used to similar degrees towards prediction of retention time. This

1 approach could be used with quantitative structure-activity relationships alone, but it
2 is recommended that analyte pK_a should be experimentally derived for ANN training.
3 Ultimately, prediction of retention times in archived screening data would simplify
4 and aid data reprocessing as a complementary tool to retrospective analysis both in the
5 identification of unknowns and where reference materials were not originally included
6 in the analytical screen. Therefore, the combination of full data capture HRMS and *in*
7 *silico* predictive approaches could improve the capability for semi-targeted urine
8 sample screening before, during and after major sporting events.

9 Acknowledgements

10 The authors would like to thank Dr David Barlow for providing constructive
11 comments on earlier drafts of this manuscript.

12 Supplementary Information

13 Tables containing the raw data relating to experimental and predicted retention times
14 for all compounds and experiments herein; molecular descriptors and associated ANN
15 input data; and all ANN models and architectures. This material is available free of
16 charge via the Internet at <http://pubs.acs.org>.

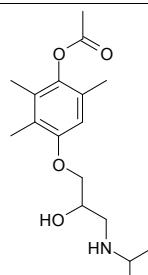
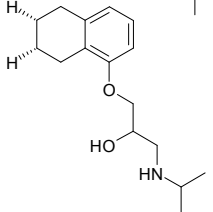
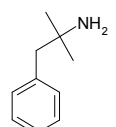
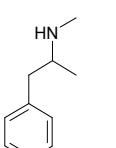
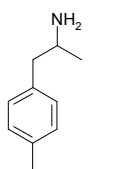
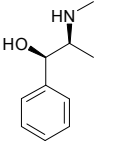
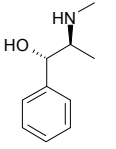
17 References

- 18 (1) World Anti-Doping Agency, The World Anti-Doping Code: The 2011
19 Prohibited List Montreal, Canada, **2012**.
- 20 (2) M. Thevis, A. Thomas, W. Schaezner, Analytical and Bioanalytical
21 Chemistry, **2011**, *401*, 405.
- 22 (3) I. Ojanpera, M. Kolmonen, A. Pelander, Analytical and Bioanalytical
23 Chemistry, **2012**, *403*, 1203.
- 24 (4) A. Musenga, D.A. Cowan, Journal of chromatography. A, **2013**, *1288*, 82.
- 25 (5) World Anti-Doping Agency, International Standard for Laboratories, Version
26 7.0, Montreal, Canada, **2013**.
- 27 (6) M. Alvarez-Guerra, D. Ballabio, J.M. Amigo, R. Bro, J.R. Viguri,
28 Environmental Pollution, **2010**, *158*, 607.
- 29

- (7) H.A. Zaqoot, A. Baloch, A.K. Ansari, M.A. Unar, *Applied Artificial Intelligence*, **2010**, *24*, 667.
- (8) I.K. Nikolos, M. Stergiadi, M.P. Papadopoulou, G.P. Karatzas, *Hydrological Processes*, **2008**, *22*, 3337.
- (9) R. Shoji, *Current Computer-Aided Drug Design*, **2005**, *1*, 65.
- (10) X.K. Wang, T.H. Liu, S.Y. Cao, *Predictions of water and sediment processes using artificial neural networks in small watershed*, **2004**, pp
- (11) P.E. Morgan, D.J. Barlow, M. Hanna-Brown, R.J. Flanagan, *Chromatographia*, **2012**, *75*, 693.
- (12) J.E. Madden, N. Avdalovic, P.R. Haddad, J. Havel, *Journal of Chromatography A*, **2001**, *910*, 173.
- (13) N.S. Quiming, N.L. Denola, S.R. Bin Samsuri, Y. Saito, K. Jinno, *Journal of Separation Science*, **2008**, *31*, 1537.
- (14) V.K. Gupta, H. Khani, B. Ahmadi-Roudi, S. Mirakhorli, E. Fereyduni, S. Agarwal, *Talanta*, **2011**, *83*, 1014.
- (15) A.A. D'Archivio, A. Giannitto, M.A. Maggi, F. Ruggieri, *Analytica Chimica Acta*, **2012**, *717*, 52.
- (16) A.A. D'Archivio, A. Incani, F. Ruggieri, *Journal of Chromatography A*, **2011**, *1218*, 8679.
- (17) A.A. D'Archivio, M.A. Maggi, F. Ruggieri, *Journal of Separation Science*, **2010**, *33*, 155.
- (18) A.A. D'Archivio, M.A. Maggi, F. Ruggieri, *Analytica Chimica Acta*, **2011**, *690*, 35.
- (19) A.A. D'Archivio, F. Ruggieri, P. Mazzeo, E. Tettamanti, *Analytica Chimica Acta*, **2007**, *593*, 140.
- (20) R. Berges, V. Sanz-Nebot, J. Barbosa, *Journal of Chromatography A*, **2000**, *869*, 27.
- (21) M. Pompe, M. Razinger, M. Novic, M. Veber, *Analytica Chimica Acta*, **1997**, *348*, 215.
- (22) N.S. Quiming, N.L. Denola, I. Ueta, Y. Saito, S. Tatematsu, K. Jinno, *Analytica Chimica Acta*, **2007**, *598*, 41.
- (23) K. Shinoda, M. Sugimoto, N. Yachie, N. Sugiyama, T. Masuda, M. Robert, T. Soga, M. Tomita, *Journal of Proteome Research*, **2006**, *5*, 3312.
- (24) K. Petritis, L.J. Kangas, B. Yan, M.E. Monroe, E.F. Strittmatter, W.-J. Qian, J.N. Adkins, R.J. Moore, Y. Xu, M.S. Lipton, D.G.C. Li, R.D. Smith, *Analytical Chemistry*, **2006**, *78*, 5026.
- (25) A. Malenovic, B. Jancic-Stojanovic, N. Kostic, D. Ivanovic, M. Medenica, *Chromatographia*, **2011**, *73*, 993.
- (26) World Anti-Doping Agency, *Technical Document – TD2010MRPL.*, Montreal Canada, **2010**.
- (27) L. Barron, J. Havel, M. Purcell, M. Szpak, B. Kelleher, B. Paull, *Analyst*, **2009**, *134*, 663.
- (28) D. Livingstone, *Theoretical property predictions*, in *Frontiers in Medicinal Chemistry*, Bentham Sci. Publ. Ltd., **2004**, pp 545.
- (29) J. Schappler, D. Guilleme, S. Rudaz, J.-L. Veuthey, *Electrophoresis*, **2008**, *29*, 11.
- (30) R.G. Devlin, K.L. Duchin, P.M. Fleiss, *British Journal of Clinical Pharmacology*, **1981**, *12*, 393.
- (31) T. Katsu, K. Ido, K. Kataoka, *Analytical Sciences*, **2001**, *17*, 745.

1 (32) Y. Fan, Y.-Q. Feng, J.-T. Zhang, S.-L. Da, M. Zhang, Journal of
2 Chromatography A, **2005**, *1074*, 9.
3 (33) P.J. Twitchett, A.E.P. Gorvin, A.C. Moffat, Journal of Chromatography **1976**,
4 *120*, 359.
5 (34) X. Liu, C. Pohl, A. Woodruff, J. Chen, Journal of Chromatography A, **2011**,
6 *1218*, 3407.
7 (35) R. Jones, G. Marnham, Journal of Pharmacy and Pharmacology, **1980**, *32*,
8 820.
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 1. Performance of ANNs for predictions of t_r for isobaric and isomeric compounds.

Compound	Monoisotopic Mass (Da) (Measured m/z for $[M+H]^+$ ion)	Molecular Formula	Structure	AlogP (MlogP)	pK_a^{ref}	t_r^E (min)	t_r^P (min)
Metipranolol	309.1940 (310.2013)	$C_{17}H_{27}NO_4$		2.86 (2.35)	9.18 ²⁹	5.62	6.02
Nadolol	309.1940 (310.2013)	$C_{17}H_{27}NO_4$		1.15 (1.36)	9.67 ³⁰	2.81	2.52
Phentermine	149.1205 (150.1277)	$C_{10}H_{15}N$		1.84 (2.55)	10.1 ³¹	2.91	2.69
Methamphetamine	149.1205 (150.1277)	$C_{10}H_{15}N$		2.07 (2.55)	9.87 ³²	2.59	2.80
<i>p</i> -methyl-amphetamine	149.1205 (150.1277)	$C_{10}H_{15}N$		2.12 (2.55)	10.0 ³³	3.45	2.75
Ephedrine							
SMILE: <chem>O[C@H](c1ccccc1)[C@@H](NC)C</chem>	165.1154 (166.1226)	$C_{10}H_{15}NO$		1.24 (1.66)	9.60 ³⁴	2.00	1.84
Pseudoephedrine							
SMILE: <chem>O[C@@H](c1ccccc1)[C@H](NC)C</chem>	165.1154 (166.1226)	$C_{10}H_{15}NO$		1.24 (1.66)	9.80 ³⁵	2.00	1.85

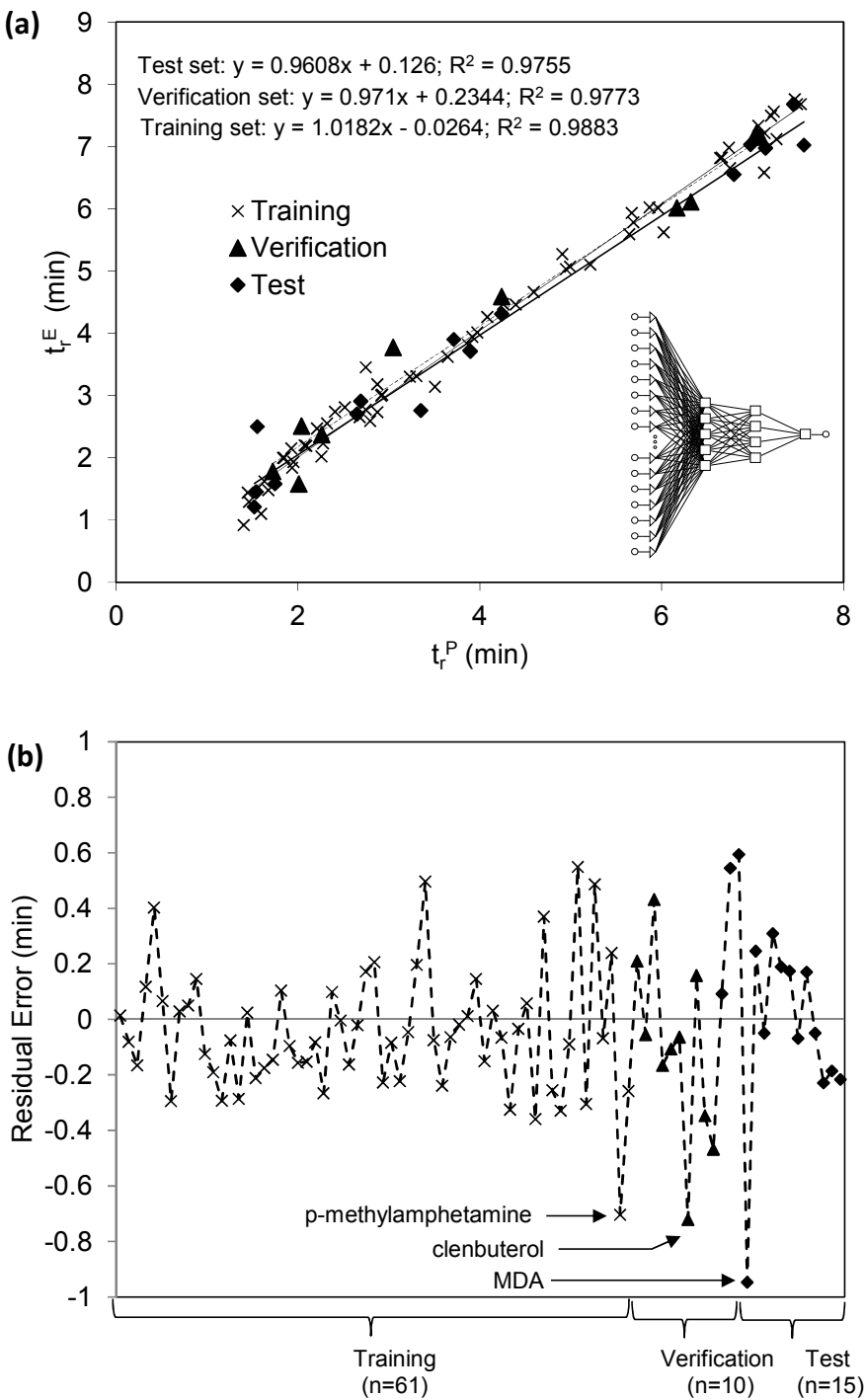


Figure 1. (a) Correlation between experimentally-derived (t_r^E) and predicted (t_r^P) retention times using the optimised 18:5:4:1 multilayer perceptron (inset) trained for 2000 epochs; and (b) residual errors in t_r^P using the optimised network for all analytes ($n = 86$). All raw data represented in Table S2 in the SI.

(SINGLE COLUMN FIGURE)

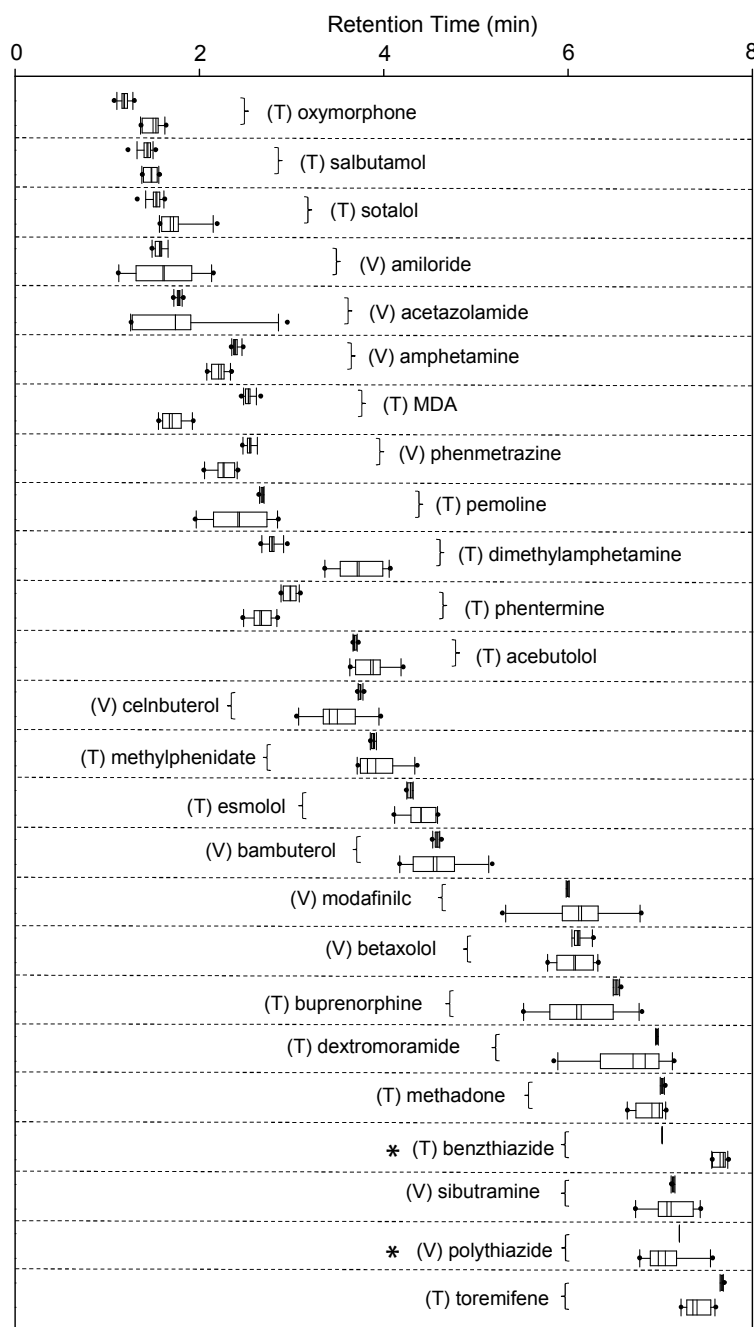


Figure 2. Paired comparisons of measured t_r^E variability (left-hand box) from $n=15$ volunteer urine data versus t_r^P variability for $n=10$ replicate 18:5:4:1 MLP networks (right-hand box). Data is presented for the verification (marked V) and blind-test compounds (marked T) only. Boxes represent 25-75th percentile, whiskers represent 10-90th percentile and dots are outliers. Thin lines represent the median and thick lines represent the mean. *Experimental variance not determined for benthiiazide and polythiazide and reported values represent a single measurement. For all raw data (including for the training set) please refer to Tables S3 & S4 in the SI.

(SINGLE COLUMN FIGURE)

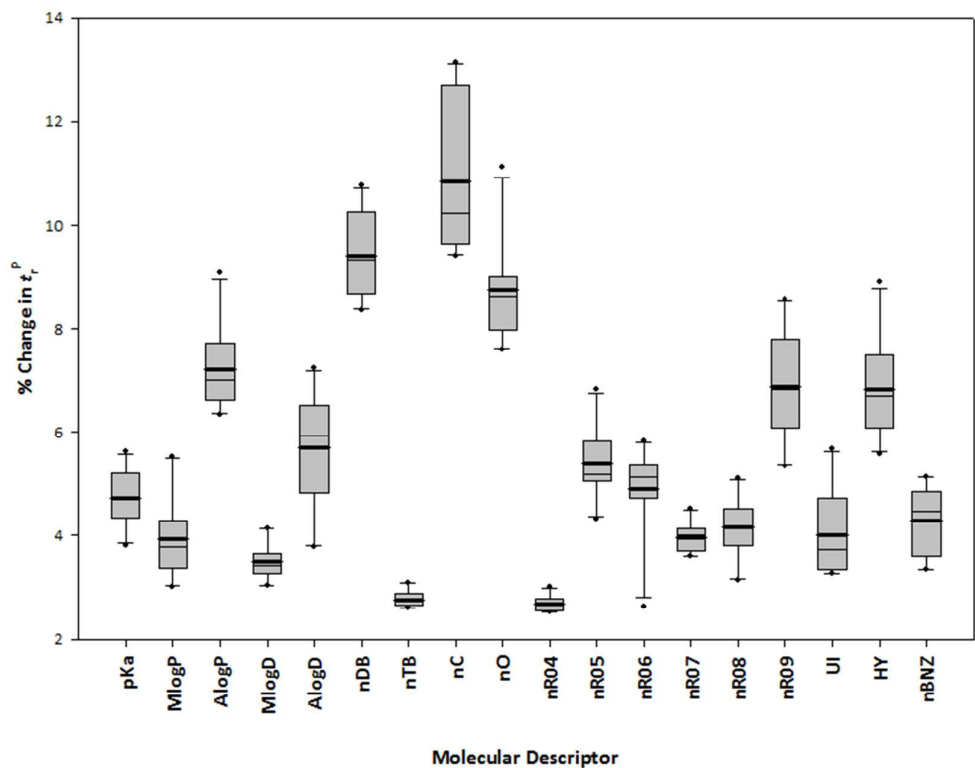


Figure 3. Percentage change in t_r^P relative to t_r^E upon systematic removal of each molecular descriptor from 10 replicate 15:5:4:1 MLP networks. Boxes include data from the 25th-75th percentile, the median (thin line), and the mean (thick line). Error bars include the 5th and 95th percentile and dots represent outliers. For all molecular descriptor definitions, please see experimental section or Table S1 in the SI.

(SINGLE COLUMN FIGURE)